

ANTUSD: A Large Chinese Sentiment Dictionary

Shih-Ming Wang and Lun-Wei Ku

NLPLab, Institute of Information Science, Academia Sinica

May 25, 2016

Outline

Motivation

Related Work

Corpora Resource

Motivation

- Sentiment dictionary
 - A building block of sentiment analysis & opinion mining
 - Applied as markers or machine learning features
- Augmented NTU¹ Sentiment Dictionary (ANTUSD)
 - Lack of Chinese resource
 - Big & complete
 - Expert labeled sentiment & machine predicted sentiment scores

¹The original authors of NTUSD were researchers at National Taiwan University

Corpora Resource I

- Words and labels were collected from several sentiment corpora (2006~2010)
- Word-base, context free
 - NTUSD
 - A widely used Chinese sentiment dictionary
 - Labels: **POS** and **NEG** (2812/8276)
 - ACIBiMA²
 - Built to test Chinese morphological structure and sentiment
 - Labels: **POS**, **NEU**, **NEG**, **NONOP**, and **NOT**
 - **NONOP** consists of regular non-emotion words
 - **NOT** consists of incorrectly segmented words

²Advanced Chinese Bi-Character Word Morphological Analyzer

Corpora Resource II

- Sentence-based, context dependent
 - NTCIR³ MOAT Task Dataset
 - Annotators labeled each sentences as **Positive**, **Neutral**, and **Neutral**
 - Annotators also labeled sentiment words as they read sentences
 - Label count for each word is it realted to its frequency
 - Chinese Opinion Treebank
 - The labeling process is similar to that of NTCIR

³<http://research.nii.ac.jp/ntcir/index-en.html>